Categoría - Comunicaciones y experiencias: Comunicaciones científicas; Experiencias profesionales.

Eje1 - Convencer de nuestro impacto: resultados y valor
Tipo de trabajo1 - Innovación como un valor de los servicios de información
Tipo de trabajo2 - Transformar ideas en proyectos o negocios

Autor
Prof. Dr. Manuel Blázquez Ochando
manuel.blazquez@pdi.ucm.es

Filiación
Universidad Complutense de Madrid
Facultad de Ciencias de la Documentación
Departamento de Biblioteconomía y Documentación

Título
Desarrollo tecnológico y documental del webcrawler Mbot: prueba de análisis web de la universidad española

Resumen:
Las investigaciones webmétricas y cibermétricas exigen el uso de herramientas específicamente diseñadas para la recolección de información en la red. Este planteamiento implica el uso de programas webcrawler que en muchos casos resultan complejos de adaptar y configurar. En consecuencia se propone la innovación y el desarrollo tecnológico de un nuevo concepto de webcrawler, denominado *Mbot*, cuya finalidad es acercar al documentalista esta tecnología y permitir al investigador implementarlo de forma rápida y efectiva. En la consecución de este objetivo se presenta el mecanismo de ejecución del programa, sus características y una prueba de análisis de la web partiendo de una semilla de enlaces constituida por 147 sedes web de universidades españolas.

Palabras clave:
Recuperación de información, webcrawler, cibermetría, automatización, tecnologías de la información, herramientas documentales

Title:
Technical and documentary development of Mbot webcrawler: web analysis test of the Spanish university

Abstract:
Cibermetrical and webometrical researches demand tools expressly designed for information harvesting in the Net. So far webcrawler applications have been used for this purpose, but most of them are very difficult to configure and to adapt for the purpose. Therefore we propose here a new concept of webcrawler, called *Mbot*, whose goal is to make this technology more adaptable to the work of the information scientist, making it at the same time quicker and more efficient. Here the execution engine of the application is introduced, its characteristics and an analysis trial of the web, starting with a link seed represented by 147 websites of Spanish universities.

Keywords:
Information retrieval, webcrawler, cybermetrics, automation, information technology, documentary tools

1. Planteamiento y objetivos

La investigación parte de la problemática de gestionar herramientas sencillas y efectivas para el análisis de la web. Se plantea la innovación y desarrollo de un programa webcrawler de fácil instalación, configuración y ejecución, que obtenga datos e información tabulada y tratada para su empleo extensivo en investigaciones cibermétricas, obteniendo información útil sobre la topografía, enlaces, documentos, contenidos y en definitiva características de la web que se pretende analizar. Seguidamente se define el pliego de especificaciones de la herramienta, la descripción de su desarrollo, aspectos conflictivos, soluciones y funcionamiento. Finalmente se presenta una prueba de análisis web de las universidades españolas, aportando su análisis de enlaces, características topográficas y de contenidos.

2. Antecedentes

Los programas webcrawler, constituyen una de las claves metodológicas y técnicas de las investigaciones cibermétricas, del desarrollo de estudios de grafo e incluso de la elaboración de mapas gráficos de la web. La principal barrera para la consecución de tales objetos es la configuración y gestión de las herramientas, que en mayor medida han sido desarrolladas con especificaciones y soportes más complejos, restringidos a sus propios desarrolladores y especialistas. Es el caso de programas como Apache Nutch o Heritrix cuyas capacidades están sobradamente reconocidas, pero cuya gestión y configuración resulta complicada. Este hecho está recogido en algunos manuales de instalación y configuración especializados como los de (BUENO LÓPEZ, J., 2010) y (SIGURDSSON, K. et al., 2007) donde esclarecen que los parámetros de control del programa deben ser modificados en su propio código fuente para definir el comportamiento del webcrawler durante el análisis de un dominio web, así como para el establecimiento de filtros de extensión, palabras clave, categorías o clasificaciones. Ello puede desembocar en procesos metodológicos muy complejos que dificultan la consecución del objeto de estudio. Este hecho no es novedoso, ya que en la literatura del sector se encuentran definidas las propiedades y características de la arquitectura de este tipo de programas (THELWALL, M., 2001, p.320) destacando que los webcrawler deberían seguir un modelo de arquitectura distribuida, capaz de operar de forma autónoma y cuyos componentes sean de fácil instalación. Un año más tarde junto con los primeros estudios extensivos de la web (SHKAPENYUK, V. and Suel, T., 2002, p.358) se identifican como problemas, los métodos de gestión y reconfiguración de las herramientas de crawling, para obtener un mayor control en los análisis especializados de la información publicada. En este sentido el estudio de (SUNIL KUMAR, M. and Neelima, P., 2011, p.12), desvela en sus conclusiones que uno de los problemas clave en el desarrollo de webcrawler es la interfaz de control e interacción entre la base de datos y el agente o algoritmo empleado en los procesos de rastreo, especialmente cuando el análisis se desarrolla sobre la web invisible.

En relación a los antecedentes del webcrawler Mbot, su desarrollo comienza en el año 2010 como resultado de la búsqueda de métodos sencillos y alternativos para el análisis de la web que solucionaran la problemática descrita. La primera versión operativa del programa se termina en el año 2011, momento a partir del cual se empiezan a realizar las primeras pruebas de funcionamiento (BLÁZQUEZ OCHANDO, M., 2011) en las que se presenta un demostrador público en línea en el que es posible reconocer cómo Mbot es capaz de extraer la información de una página web, depurar su contenido textual, metadatos, enlaces y documentos vinculados, presentándolos sistemáticamente a modo de informe. Posteriormente, se realizaron los primeros análisis en profundidad, orientados a la usabilidad y caracterización de los sitios web de las agencias espaciales americana y europea (BLÁZQUEZ OCHANDO, M. and Serrano Mascaraque, E., 2011), obteniendo una experiencia muy valiosa para la mejora y desarrollo de una versión definitiva, que tuviera en cuenta una mayor variedad de excepciones, casos y errores frecuentes. Actualmente Mbot se encuentra en la fase final de desarrollo, en la que se están perfeccionando y ampliando sus capacidades originales, para dar cabida al análisis automático del grafo, estadísticas del análisis de enlaces, contenidos y el aprovechamiento de los recursos como herramienta de minería de datos y reaprovechamiento de la información.

3. Desarrollo tecnológico y documental de Mbot

La idea básica para el desarrollo de un webcrawler es la concreción de los datos, informaciones o contenidos que se pretenden recuperar en la web, esto es definir qué elementos están presentes en las diversas páginas que el programa analizará, destacando esencialmente los enlaces, metadatos, meta-etiquetas, canales de sindicación, imágenes, documentos, archivos multimedia, código fuente y texto completo depurado. Pero la información debe ser obtenida a partir de una muestra básica de enlaces relativos al dominio o área de la

web que se pretenda analizar, también denominada semilla. A partir de este punto el programa debe extraer el código fuente de la página correspondiente a cada enlace de la muestra, detectar si existen enlaces derivados y registrarlos para continuar su análisis posteriormente, repitiendo el proceso tantas veces como sea indicado en su configuración. A cada repetición de este proceso recursivo se le denomina nivel de análisis. Esto quiere decir que el primer nivel contendrá la información de las páginas web correspondientes a los enlaces de la semilla, el segundo nivel lo conformarán las páginas enlazadas al nivel anterior y así sucesivamente conforme se amplíe la profundidad del análisis. Añadido a esto, un webcrawler debe permitir el empleo de diversos filtros para restringir el análisis según el servidor, dominio o página web especificada, así como la capacidad para rastrear extensiones o enlaces que cumplan condiciones establecidas por expresiones regulares. Esto es la capacidad para restringir y perfilar adecuadamente los análisis web y reducir en tal caso el tiempo de ejecución de la operación.

Otro aspecto fundamental en el pliego de especificaciones es el método de almacenamiento de la información. Dependiendo de la investigación que se lleve a cabo, la información debe poder ser almacenada en tablas diferenciadas para cada nivel de profundidad del análisis, en una tabla única o en archivos de texto para su procesamiento con terceros programas. La diferencia entre tales métodos de almacenamiento viene determinada por el volumen de información a tratar y la jerarquización de los análisis cibermétricos. De hecho puede llegar a ser necesario comprobar la variabilidad de la topografía y su caracterización en cada nivel de análisis (para lo que se recomienda el almacenamiento diferenciado de la información de tales niveles en tablas independientes) o la elaboración de estadísticas cuantitativas (en la que el almacenamiento en tabla única o archivo de texto es más adecuado por su simplicidad).

En cuanto al apartado de funcionamiento, el webcrawler debe ser fácilmente configurable en cuanto a su comportamiento y ejecución. Esto es la definición de estrategias de análisis, optimizadas para los distintos usos que puedan hacerse del mismo, por ejemplo el análisis de la web orientado al establecimiento de un ranking (en el que debe calcularse el pagerank y la relevancia de los documentos), el análisis orientado a la usabilidad y accesibilidad (donde se detecten excepciones, errores y códigos mal programados), el análisis web de los contenidos (orientado a la minería de datos, extracción de textos, contenidos y documentos) y el análisis de enlaces (para el que la interrelación de los enlaces entrantes y salientes, así como la definición del grafo es el objetivo principal). Para todos los casos descritos el webcrawler tiene unas rutinas de funcionamiento pre-programadas que permiten definir cuál será su modo de trabajo en cada momento. Una vez delimitado el objeto, funciones y procesos que deberá llevar a cabo el programa Mbot, la idea comienza a tomar forma, partiendo de un esquema básico de funcionamiento, véanse *figuras 1, 2, 3 y 4*.

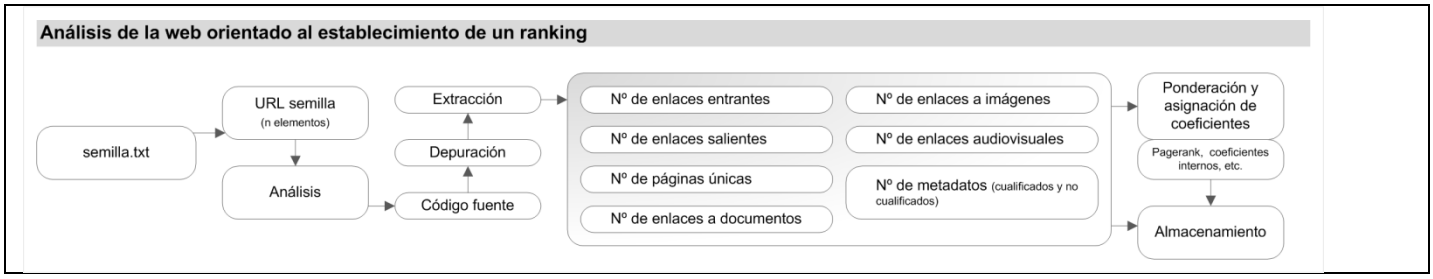


Figura 1. Borrador del proceso de análisis web para rankings

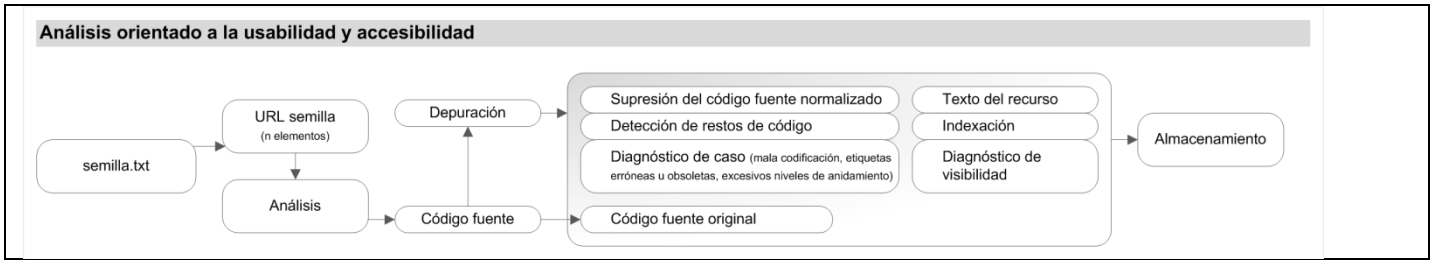


Figura 2. Borrador del proceso de análisis para evaluación de la usabilidad y accesibilidad web

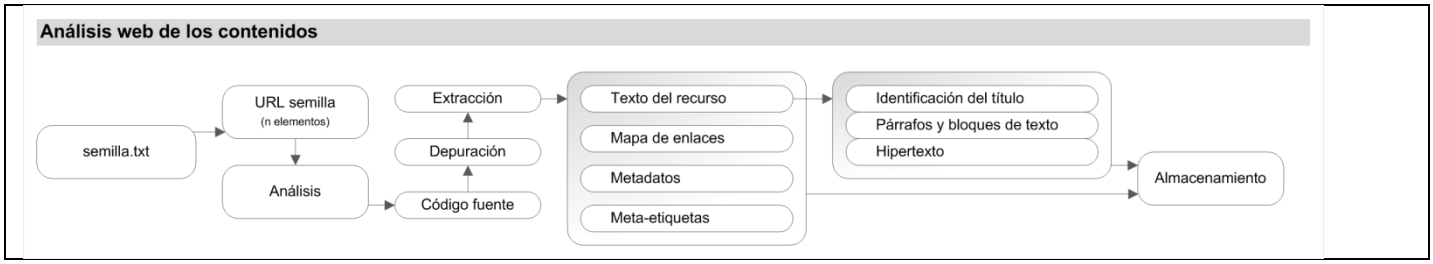


Figura 3. Borrador del proceso de análisis de contenidos web

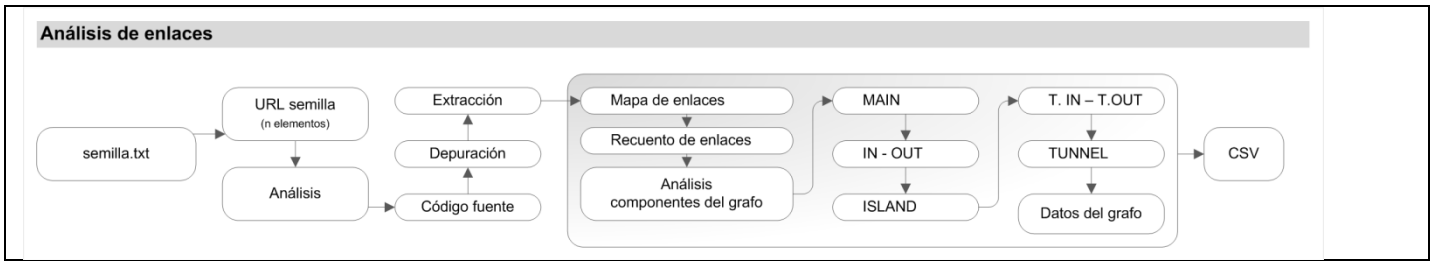


Figura 4. Borrador del proceso de análisis de enlaces web

Identificado y plasmado el programa sobre un esquema borrador, se debe definir la arquitectura del mismo. En este caso se desea crear un webcrawler de tipo distribuido y fácil instalación en cualquier servidor web. Para ello se eligió la plataforma Apache, MySQL y PHP (AMP). Con estas bases, se establece la estructura de las tablas contenedoras de la información que se obtendrá del rastreo y análisis de la web, véase *tabla1*.

Campo	Descripción
id	Identificador auto-numérico.
url1	El campo *url1* identifica la URL principal objeto de análisis siempre que el campo *url2* quede vacío. Si *url2* consta de contenido, el campo *url1* representa la dirección de procedencia.
url2	El campo *url2* identifica la URL principal objeto de análisis siempre que el campo *nlevel* indique un valor superior a *N1*; por ejemplo *N2*, *N3*, *N4*…
date	Fecha y hora de registro en formato ISO 8601.
updated	Fecha de actualización del registro.
nlevel	Nivel de profundidad del análisis. Por defecto el sistema se ha diseñado para contemplar hasta 10 niveles.
rankmbot	Valor de ranking interno.
rankgoogle	Valor de pagerank de Google.
title	Título de la página web.
metatag	Contenido de las Meta-etiquetas HTML <meta> no identificadas como Dublin Core.
metadata	Metadatos Dublin Core básicos (prefijo DC) y refinamientos (prefijo DCTERMS).
linkmap	Recopilación de todos los enlaces de la página web analizada, separados por barras verticales (|) para facilitar su manipulación como arrays o matrices de datos.
linksync	Recopilación de enlaces a canales de sindicación RSS, Atom y RDF 1.0, separados por barras verticales (|)
linkdoc	Recopilación de todos los documentos en formatos pdf, doc, rtf, xls, ppt, etc. separados por barras verticales (|)
linkimg	Enlaces a imágenes, separadas por barras verticales (|)
linkaud	Enlaces de archivos de audio, separadas por barras verticales (|)
linkvid	Enlaces de archivos de vídeo, separadas por barras verticales (|)

sourcecode	Código fuente completo de la página analizada. Utilizado para detectar errores de usabilidad, codificación, accesibilidad etc.
indexer	Texto completo indexado de la página web.

Tabla 1. Estructura básica de campos de las tablas de contenidos del webcrawler

Este método del almacenamiento de la información, permite estructurar los datos recuperados y acceder a ellos de forma sencilla y efectiva, así como determinar fácilmente en qué nivel fueron analizados los enlaces y de qué páginas fueron extraídos. Como es lógico el análisis de los co-enlaces (ORTEGA, J.L. and Aguillo, I.F., 2007, p.272), la detección de duplicaciones (FABA PÉREZ, C. et al., 2004, p.109) y los elementos del grafo (GRAELLS, E. and Baeza Yates, R., 2007, p.6), pueden ser estudiados a partir de la información almacenada en la base de datos, mediante el cumplimiento de las condiciones establecidas para cada componente, véase *tabla2*.

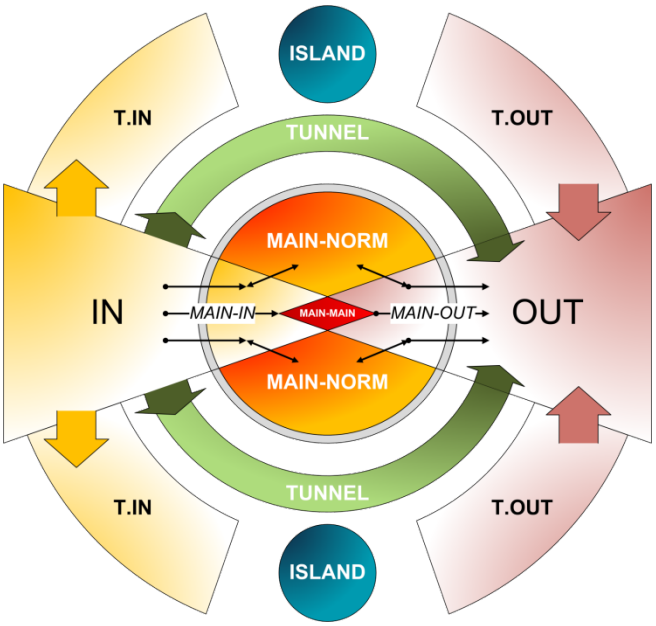
Estructura del grafo	Componentes	
	MAIN	url1←→ url main ←→ url2 linkmap url ←→ url main ←→ linkmap url
	IN	url1 → url main url2 → url main linkmap url → url main
	OUT	url main → url1 url main → url2 url main → linkmap url
	ISLAND	→ url1 ← → url2 ← → linkmap url ←
	T.IN	(url1 → url main) → linkmap url (url2 → url main) → linkmap url
	T.OUT	(url main → url1) → linkmap url (url main → url2) → linkmap url
	TUNNEL	url1 → url2 url2 → url1 linkmap url → linkmap url

Tabla 2. Componentes del grafo de mbot

En la presente tabla se observan tales condiciones, para distinguir los componentes de un grafo en un análisis cibermétrico. Este problema fue resuelto estableciendo el siguiente orden en la obtención de los distintos componentes (Main, in, out, island, tentacle in, tentacle out y tunnel). En segundo lugar se diseñó un algoritmo para determinar qué porcentaje de enlaces estaban totalmente interconectados, cuáles enlazaban solamente a los principales, cuáles eran enlazados desde los principales, cuáles estaban aislados o directamente vinculaban los enlaces entrantes con los salientes. Esta contabilización automática da como resultado una información clave para determinar qué páginas web requieren una mayor vinculación o simplemente cómo evoluciona un determinado área de conocimiento en la web, pudiendo observar de forma consecutiva el crecimiento de la misma.

En cuanto al funcionamiento de Mbot, *véase figura 5*, el proceso comienza cuando se activa el primer nivel de análisis de los diez disponibles en la actual versión del programa. El arranque del webcrawler se efectúa mediante la carga de la semilla editada con anterioridad. Se extraen todos los enlaces contenidos, efectuando su contabilización como mecanismo de control. Cada enlace es depurado para eliminar cualquier espacio o sustituir cualquier carácter especial que dificulte o imposibilite la carga de la página web. A continuación se

comprueba la conexión por socket o lo que es lo mismo, la existencia o no de la página web referida en el enlace. Este paso es fundamental para garantizar un correcto funcionamiento del sistema, ya que de no hacerlo, pueden producirse bloqueos en la ejecución del programa, derivados del intento continuo de procesar una página o contenido inexistente. Una vez verificada la página, se imprime en pantalla un icono correspondiente al enlace analizado, que servirá para generar una idea del progreso del análisis. Acto seguido se ejecuta una rutina de programación basada en la librería cURL (cURL installation, 2012), de forma tal que procesa el enlace, aplicando los ajustes de configuración anteriormente establecidos. El resultado de este proceso es la obtención del código fuente completo de la página web que es almacenada temporalmente para su tratamiento.

Para poder manipular la información contenida en el código fuente de la página web se utiliza la librería DOM (Document Object Model), ya que genera automáticamente un esquema estructural de todos sus nodos y etiquetas. Concretamente se crea un elemento DOM (The DOMElement class, 2012) a través del cual se consultan los distintos elementos de interés para el documentalista, es decir, todos los enlaces disponibles de la página, los canales de sindicación, sus metadatos e imágenes correspondientes. Dichas consultas son planteadas mediante el lenguaje XPath (SimpleXMLElement:xpath, 2012), tal y como se muestra en la *figura5*. La información que se extrae es almacenada en matrices de datos, para posteriormente segmentarla en una cadena de texto cuyos valores son separados por un carácter especial, en este caso la barra vertical (|). Este procedimiento permite una fácil reutilización de la información separada por este carácter, evitando en todo caso mezclar los datos obtenidos.

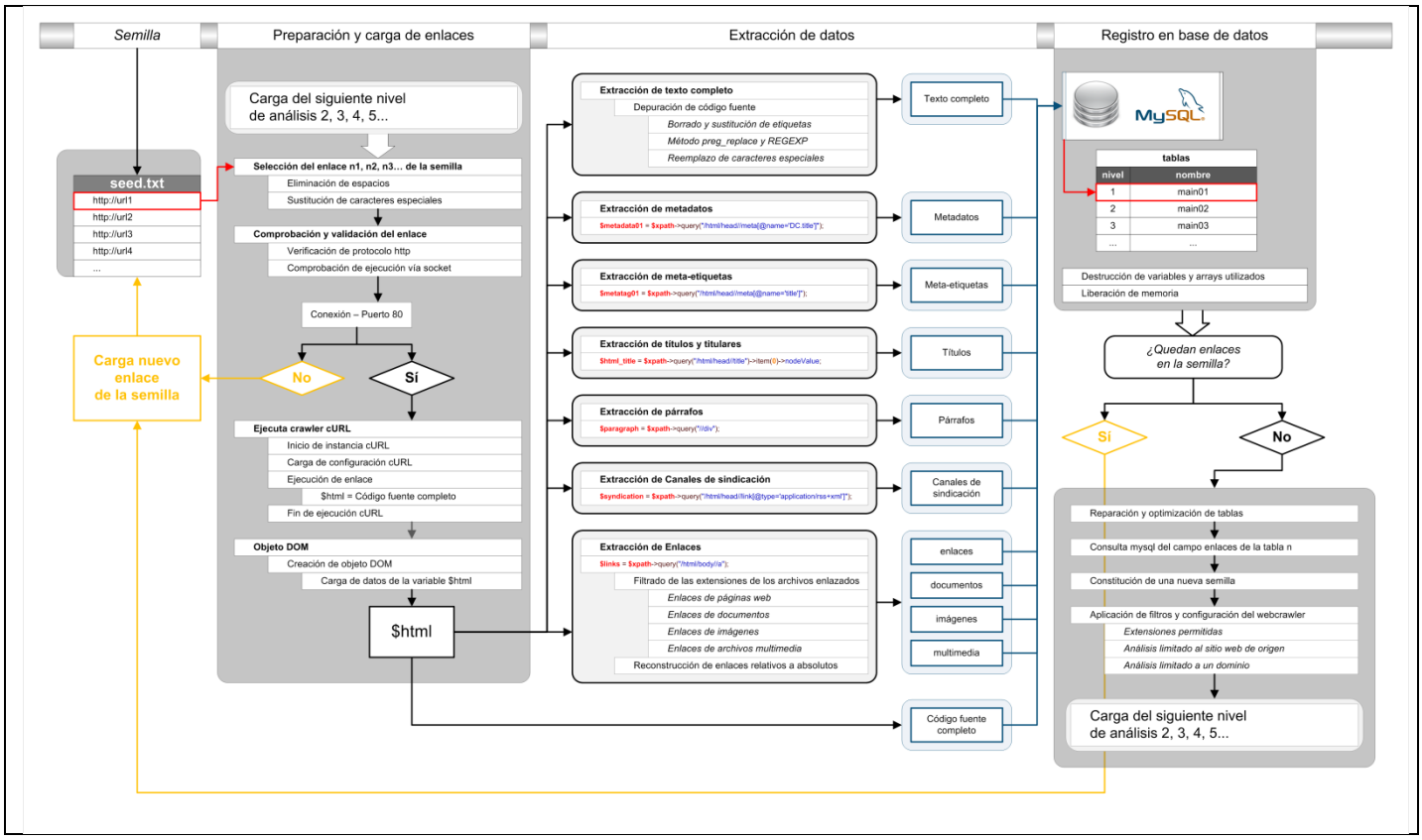


Figura 5. Esquema del procesamiento de la web y la extracción de sus datos mediante XPath

El siguiente paso es la preparación del texto disponible en la página. Éste se encuentra entreverado con el etiquetado HTML, propio del código fuente que se ha extraído anteriormente. Para ello se lleva a cabo un proceso de limpieza y depuración del etiquetado, así como el escapado y reemplazo de caracteres especiales, comillas y demás signos que puedan perjudicar la labor de indexado en la base de datos. El resultado del proceso no siempre es perfecto, puesto que en algunos casos quedan residuos del código fuente original. A pesar de ello se logra captar todo el contenido textual de la página que facilitará posteriormente el tratamiento y recuperación dentro del conjunto de la muestra.

Una vez se dispone de toda la información, se procede al registro en base datos. Tal proceso es muy delicado ya que el motor de almacenamiento MyISAM de MySQL no soporta código fuente cuyo etiquetado esté mal

cerrado o escapado. Cuando los datos de la página son almacenados, se destruyen todos los arrays y variables que contengan información del proceso. Este método de actuación asegura que no existirá información residual de una página anterior, garantizando que todos los datos almacenados sean únicos y exclusivos de cada enlace en cada momento.

Otra operación de relevancia es la reparación de todos los enlaces almacenados en la tabla de la base de datos. Esto es debido a que una parte de los enlaces recopilados o no son válidos, o constituyen enlaces relativos, o conforman conjuntos vacíos, o constituyen enlaces a documentos. Todos estos casos son analizados utilizando expresiones regulares, eliminando aquellos enlaces no válidos, reconstruyendo enlaces relativos en absolutos, eliminado los conjuntos vacíos, y almacenando separadamente los enlaces a documentos, para ser actualizados en cada registro de la base de datos. Es en este punto, donde también se aplica el filtrado por extensiones, ya que se opera directamente sobre los recursos que se emplearán en sucesivos niveles de análisis.

4. Prueba de funcionamiento: análisis web de la universidad española

El análisis de la web universitaria española ha sido abordado anteriormente en diversos estudios con gran rigurosidad científica. Por ejemplo, (THELWALL, M. and Aguillo, I.F., 2003, p.296) efectúan un análisis de las 64 sedes universitarias españolas, su tamaño, cantidad de enlaces a páginas y sitios, interrelación de enlaces entrantes y salientes, tipología de la web o uniformidad de la misma. Un año más tarde el trabajo de (AGUILLO, I.F. et al., 2004) recoge un análisis más completo de 4000 sedes web universitarias, centrado en la identificación de los dominios universitarios utilizando motores de búsqueda comerciales, detectando el tamaño de la web académica, su distribución de subdominios, tipología de contenidos según formatos, visibilidad y factor de impacto. Finalmente, el trabajo de (PINTO MOLINA, M. et al., 2004) tiene un enfoque más cualitativo en el que se toman 19 sedes universitarias españolas para determinar cuáles son sus indicadores de calidad y visibilidad de la investigación a través de sus páginas web.

Teniendo en cuenta los precedentes, la prueba del webcrawler Mbot se ha realizada sobre una muestra más reducida, de 147 sedes universitarias, recogidas en el ranking web de universidades españolas (CYBERMETRICS LAB, 2012). La profundidad del análisis ha sido limitada a 3 niveles, con objeto de reducir el tiempo de procesamiento de la información, permitiendo efectuar en tal caso varias pruebas de ejecución, reproduciendo por duplicado, verificando y validando todos los datos recopilados por el sistema. El objetivo de la prueba es demostrar la capacidad de la herramienta para obtener datos sobre la topología de la web, sus contenidos, dimensiones y características. Los datos generales del análisis pueden resumirse en la *tabla3*, con un total de 37.051 páginas analizadas para los 3 niveles de análisis, lo que supone una compleción del 90% de todos los enlaces únicos detectados en el nivel 2. El número total de enlaces únicos obtenidos es cercano a los 270.000, distribuidos en una ratio media correspondiente a 9,4 páginas por sitio web. El tamaño total de la información alojada en la base de datos supera el 1 GB de espacio en disco.

Nivel de profundidad del análisis	**Nº de recursos analizados**	Nº de enlaces únicos	Sitios web	Páginas web	Tamaño en MB
Nivel 1	**147**	8.108	1.544	6.564	5
Nivel 2	**6.479**	41.031	5.165	35.866	109.7
Nivel 3	**30.425**	220.794	19.365	201.429	920.2
Total	**37.051**	269.933	26.074	243.859	1034.9
Tiempos de ejecución	Inicio 2012-10-12T12:30:36+01:00 - Fin 2012-10-13T18:11:10+01:00				

Tabla 3. Datos generales del análisis

En relación a los documentos de texto y ofimática, los formatos predominantes son PDF (9,99%) seguido de DOC (0,63% absoluto) respecto al total de enlaces únicos extraídos, véase *tabla4*.

Formato	doc	ppt	xls	rtf	mdb	pdf
Nº de enlaces	1.171	59	124	27	4	18.479

Tabla 4. Documentos de texto y ofimática

En cuanto a los archivos audiovisuales, se descubre que no tienen una representación notable en la web universitaria española, ya que en ningún caso los formatos superan el centenar de archivos de su clase, véase *tabla5*.

Formato	swf	mp4	wmv	mp3	ogg	wma
Nº de enlaces	31	29	26	76	2	2

Tabla 5. Documentos de vídeo

En cambio, los archivos de imagen tienen una gran representación, respecto al total de enlaces únicos, destacando los formatos JPG (14,21%) y PNG (4,94%) destinados a la ilustración de la información y de las páginas web, y el formato GIF (5,00%) orientado más al interfaz gráfico, véase *tabla6*.

Formato	png	jpg	gif	tif	svg	bmp
Nº de enlaces	9130	26.282	9.258	22	0	135

Tabla 6. Distribución de los archivos de imagen

Sobre la tipología de las páginas web y su programación, se detecta que cerca de la mitad de las páginas web localizadas son de tipo estático HTML y derivados (49,47% relativo – 32,04% absoluto), y la otra mitad es web dinámica (50,53% relativo – 32,82% absoluto) en PHP y ASP, véase *tabla7*.

Formato	html	xml	rdf	rss	php	asp
Nº de enlaces	59.267	166	0	2.811	38.512	22.202
Porcentaje relativo a formatos web	Enlaces a páginas estáticas 49,47%			Enlaces a páginas dinámicas 50,53%		

Tabla 7. Tipología de páginas web según su extensión

También se ha abordado el estudio de algunas de las extensiones de dominios más recurrentes, véase tabla8, en la que se observa que, cerca de la mitad de los enlaces únicos extraídos por el webcrawler son dominios de tipo ES (45,21%), seguido de EDU (29,60%) y COM (14,97%).

Extensiones de dominios	es	fr	de	it	uk	us	com	org	net	edu	gov
Nº de sitios	3.645	84	73	53	146	7	2.838	1.207	335	1.075	19
Nº de páginas	118.385	401	819	113	686	4.438	37.573	4.965	6.785	78.814	77
Porcentaje de enlaces únicos	45,21%	0,18%	0,33%	0,06%	0,31%	1,65%	14,97%	2,29%	2,64%	29,60%	0,04%

Tabla 8. Distribución de extensiones por sitios y páginas

La macroestructura de la web universitaria española, sobre los 37.051 recursos analizados, detecta que el componente de enlaces fuertemente conexos representa el 9,60% del total, que en suma junto con la componente IN, OUT, T.IN, T.OUT y TUNNEL, representan un 15,56% de sitios web con más de 1 página. La componente ISLAND, más desconectada supone un importante porcentaje del 84,41% de los enlaces, lo que indica que el acceso a tales contenidos, no está lo suficientemente interconectado.

Componente	Porcentaje
MAIN	9,60%
IN	0,24%
OUT	2,84%

ISLAND	84,41%
TENTACLE IN	0,12%
TENTACLE OUT	2,55%
TUNNEL	0,21%

Tabla 9. Macroestructura de la web de universidades españolas

Las universidades con más enlaces únicos desde su portal son las que figuran en la *tabla10*, entre los que se observa diferencias en el número de sitios y páginas. Ello significa que no siempre las universidades con más enlaces únicos son las que más páginas tienen. También se debe tener en cuenta que el análisis a 3 niveles efectuado no recorre toda la jerarquía de páginas que pudieran existir, lo cual requeriría de un análisis más pormenorizado para destacar que otras universidades de mayores dimensiones también tienen una representación en los primeros puestos de la tabla.

Pos.	Universidad	URL	**Nº de enlaces únicos**	Nº de sitios	Nº de páginas
1	IE Business School	http://www.ie.edu/	**44.502**	3.341	41.161
2	U. de Alicante	http://www.ua.es/	**48.225**	1.915	46.310
3	IESE Business School U. Navarra	http://www.iese.edu/	**6.174**	1.447	4727
4	U. Politécnica de Cataluña	http://www.upc.edu/	**8.760**	764	7996
5	U. de Málaga	http://www.uma.es/	**2.528**	754	1774
6	U. de Jaén	http://www.ujaen.es/	**4.579**	571	4008
7	E. S. Archivística U.A. Barcelona	http://www.esaged.com/	**582**	541	41
Total de universidades			**222.247**		

Tabla 10. Ranking de universidades con más enlaces únicos en su portal

Las universidades con más documentos en sus portales web son las reseñadas en la *tabla11*. Coinciden algunos de los resultados con los de la tabla anterior, como por ejemplo la Universidad de Alicante, el IE Business School ó la Universidad Politécnica de Cataluña. Además de las razones aducidas al análisis de niveles, hay que añadir un factor que puede incidir en una peor extracción de la información, como por ejemplo el tipo de CMS utilizado por cada universidad, indicio que permitiría detectar qué sistemas de gestión de contenidos hacen posible una mejor indexación de la información.

Pos.	Universidad	URL	**Nº de enlaces a documentos únicos**
1	U. de Alicante	http://www.ua.es/	**10.998**
2	IE Business School	http://www.ie.edu/	**6.978**
3	U. de Girona	http://www.udg.edu/	**2.646**
4	U. Miguel Herández	http://www.umh.es/	**2.604**
5	U. Málaga	http://www.uma.es/	**2.228**
6	U. Politécnica de Cataluña	http://www.upc.edu/	**1.895**
7	IESE Business School U. Navarra	http://www.iese.edu/	**1.886**
Total de universidades			**66.355**

Tabla 11. Ranking de universidades con más documentos en su portal

Finalmente, el webcrawler Mbot permite generar un archivo DOT para la exportación del mapa de enlaces de la web analizada. El método para generar el gráfico de la web universitaria española es similar al expuesto en los trabajos de (MEDRANO, J.F. et al., 2011, p.4) y (RODRÍGUEZ MIRANDA, A. and Valle Melón, J.M., 2012, pp.2-4), empleando para ello el programa Graphviz. El resultado de la visualización es el que sigue en la *figura6*.

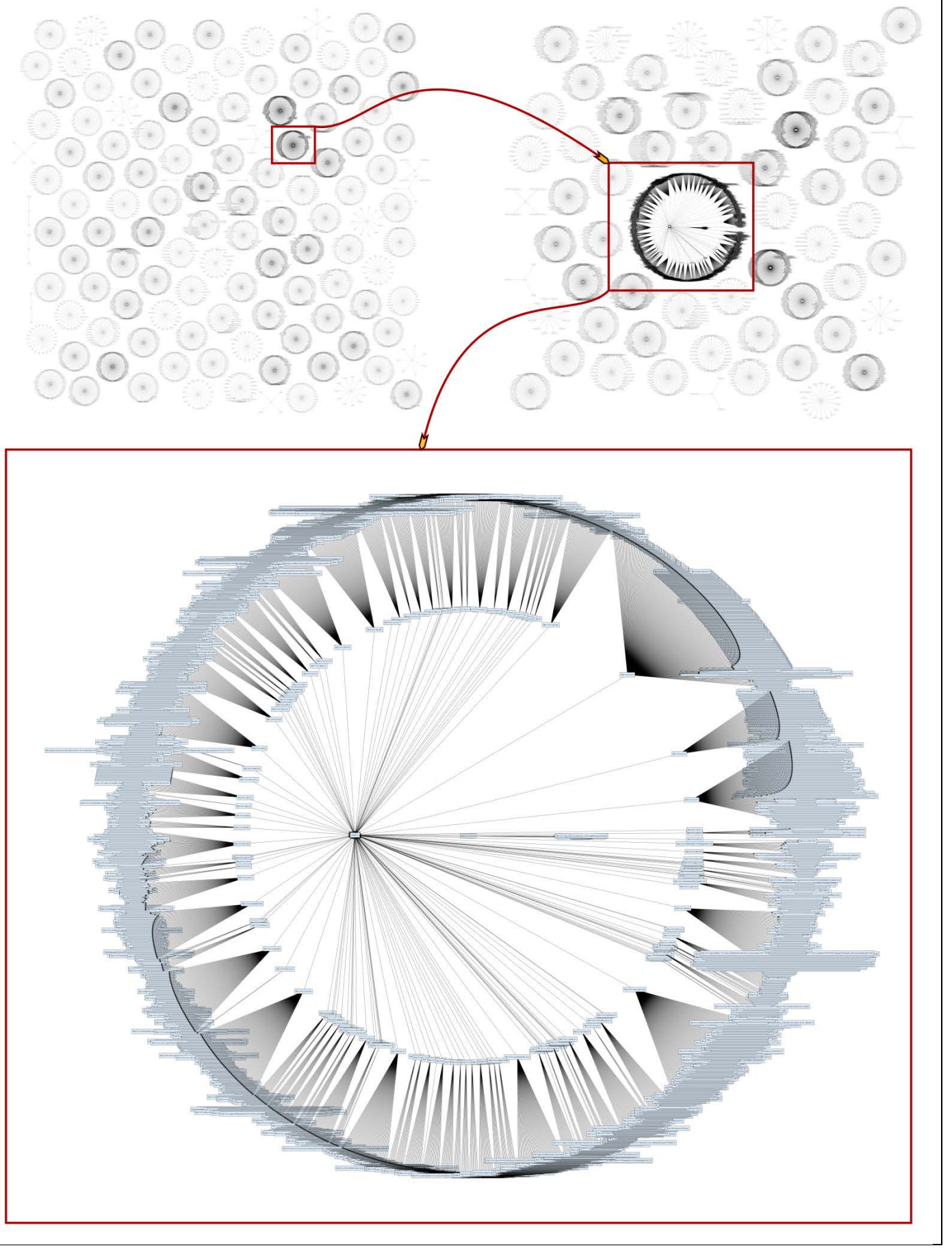


Figura 6. Representación gráfica de la web de la universidad española

5. Conclusiones y perspectiva

1. El programa webcrawler Mbot logra efectuar análisis webmétricos y cibermétricos, que incluyen el análisis básico de los componentes de la macroestrutura de la web, su composición, características y formatos de forma sencilla y completamente automática sin necesidad de efectuar posteriores procesos de tabulación.

2. Se demuestra la necesidad de seguir mejorando la efectividad del programa, ya que analiza el 90% del total de enlaces únicos que se recopilan. Este efecto puede ser debido a fallos en la composición de los enlaces relativos más complejos.

3. La herramienta hace posible los micro-estudios de la web, lo que ayudará a democratizar y ampliar el uso de los webcrawlers para todos los profesionales de la información, sin un conocimiento exhaustivo en las metrías, proporcionando información ya depurada y procesada, con la que conocer mejor el crecimiento de la web por sectores ó áreas de conocimiento.

4. El futuro de Mbot, será la consolidación del interfaz gráfico de control, de los distintos modos de ejecución a saber: 1) análisis de rankings, 2) análisis de enlaces, 3) análisis de contenidos, 4) análisis de usabilidad y accesibilidad. Finalmente, una vez superadas las pruebas de ejecución y rendimiento, la herramienta será puesta a disposición de cualquier investigador como programa de código abierto.

5. En relación a la web universitaria española, se concluye que un análisis a 3 niveles, permite obtener más de 250.000 enlaces únicos de los que 26.000 son sitios web, suponiendo en conjunto un total de más de 1GB de tamaño total.

6. Los dominios de tipo ES son los más abundantes con el 45% del total, junto con los de tipo académico EDU con un 29%. La presencia de dominios de terceros países es muy reducida, lo cual indica una baja interrelación de la web universitaria española con instituciones académicas extranjeras.

7. Los documentos de texto, imágenes y audiovisuales representan un 35% del total de enlaces únicos extraídos, quedando un 75% de sitios y páginas web entre las que se encuentra un equilibrio cuantitativo próximo al 50% entre el número de webs estáticas y dinámicas.

8. Sólo el 15% de la web universitaria española se encuentra en los valores altos de interconexión de sus enlaces. El resto queda débilmente vinculado, lo cual sugiere una red de universidades poco cooperativa. La representación gráfica de la web de la universidad española, permite corroborar este hecho, destacando el alto número de islas, comprobando cómo la web universitaria española es en sí misma una gran isla.

6. Referencias

- AGUILLO, I.F., M.B. GRANADINO GOENECHEA, C. RONDA LAÍN et al. 2004. *Factor de impacto y visibilidad de 4000 sedes web universitarias españolas. (Proyecto Estudios y Análisis 2004 EA2004-0020).*

- BLÁZQUEZ OCHANDO, M. 2011. *Primeras pruebas del mbot webcrawler*. [online]. [Consultado 2 Oct 2012]. Disponible en: http://www.mblazquez.es/documents/articulo-pruebas1-mbot.html

- BLÁZQUEZ OCHANDO, M. and E. SERRANO MASCARAQUE. 2011. Análisis de la web y usabilidad: prueba de funcionamiento de Mbot webcrawler. *In*: *X Congreso ISKO Capítulo Español (Ferrol, 30 junio - 1 julio 2011)*. Ferrol: ISKO.

- BUENO LÓPEZ, J. 2010. *Nutch: technological wiki*. [online]. [Consultado 07 Mar 2011]. Disponible en: http://thewiki4opentech.org/index.php/Nutch

- CYBERMETRICS LAB. 2012. *Ranking Web of Universities: spain*. [online]. [Consultado 2 Oct 2012]. Disponible en: http://www.webometrics.info/en/Europe/Spain

- FABA PÉREZ, C., V.P GUERRERO BOTE, and F. MOYA ANEGÓN. 2004. *Fundamentos y técnicas cibermétricas*. Badajoz: Consejería de Educación, Ciencia y Tecnología. Junta de Extremadura.

- GRAELLS, E. and R. BAEZA YATES. 2007. *Características de la Web Chilena 2007*. Santiago de Chile.

- MEDRANO, J.F., J.L. ALONSO BERROCAL, and C.G. FIGUEROLA. 2011. *Visualización de Grafos Web*. [online]. [Consultado 2 Oct 2012]. Disponible en: http://www.academia.edu/942300/Visualizacion_de_Grafos_Web

- ORTEGA, J.L. and I.F. AGUILLO. 2007. Análisis de co-enlaces: una aproximación teórica. *El Profesional de la Información*. **15**(4), pp.270-277.

- PHP GROUP. 2012. *cURL installation*. [online]. [Consultado 28 Oct 2012]. Disponible en: http://php.net/manual/es/book.curl.php

- PHP GROUP. 2012. *SimpleXMLElement:xpath*. [online]. [Consultado 28 Oct 2012]. Disponible en: http://php.net/manual/es/simplexmlelement.xpath.php

- PHP GROUP. 2012. *The DOMElement class*. [online]. [Consultado 28 Oct 2012]. Disponible en: http://php.net/manual/en/class.domelement.php

- PINTO MOLINA, M., J.L. ALONSO BERROCAL, J.A. CORDÓN GARCÍA et al. 2004. Análisis cualitativo de la visibilidad de la investigación de las universidades españolas a través de sus páginas web. *Revista Española de Documentación Científica*. **27**(3), pp.345-370.

- RODRÍGUEZ MIRANDA, A. and J.M. VALLE MELÓN. 2012. *Software to generate graphs in DOT format (v. 1.0)*. [online]. [Consultado 2 Oct 2012]. Disponible en: https://addi.ehu.es/bitstream/10810/6169/6/ldgp_sof007_grafos.pdf

- SHKAPENYUK, V. and T. SUEL. 2002. Design and Implementation of a High-Performance Distributed Web Crawler. *In*: *Proceedings. 18th International Conference on Data Engineering*. Nueva York, pp.357-368.

- SIGURDSSON, K., M. STACK, and I. RANITOVIC. 2007. *Heritrix User Manual*. [online]. [Consultado 7 Mar 2011]. Disponible en: https://pacer.ischool.utexas.edu/bitstream/2081/1708/1/user_manual.html

- SUNIL KUMAR, M. and P. NEELIMA. 2011. Design and Implementation of Scalable, Fully Distributed Web Crawler for a Web Search Engine. *International Journal of Computer Applications*. **15**(7), pp.8-13.

- THELWALL, M. 2001. A web crawler design for data mining. *Journal of Information Science*. **27**(5), pp.319-325.

- THELWALL, M. and I.F. AGUILLO. 2003. La salud de las web universitarias españolas. *Revista Española de Documentación Científica*. **26**(3), pp.291-305.